SYSTEM AND METHODS FOR PATIENT DATA AND TREATMENT MANAGEMENT

FIELD OF INVENTION

Embodiments of the present invention are directed to methods and systems for patient data and treatment management and analysis.

BACKGROUND OF THE INVENTION

In any field of medicine, there is an ever-present interest in diagnosing a patient's condition as accurately as possible. There is a further interest in establishing the patient's prognosis and implementing the most effective therapeutic treatment or treatments, especially in those instances where a variety of treatment options are available (including the option of not administering any therapy at all). It is often difficult to assess at the outset of treatment for a particular condition what therapeutic regimen will be most effective in a patient. Physicians may have to simply attempt treatment with a first therapy and later implement a second therapy if that first therapy does not effect the desired physiological response. Moreover, a patient's disease condition is dynamic in nature; it changes with time in response to treatment and as pathology progresses. This progression is generally monitored and treatment is adjusted accordingly. Yet it may be difficult to determine the most effective therapeutic treatment at various stages of clinical intervention.

Modern trends in medical research highlight the importance of understanding genetic or other physiological roots responsible for creating or facilitating the development of disease conditions. Objects of this research are to discover why two people similar in both intrinsic characteristics (*e.g.*, age, weight, sex, height, body type, family history, other genetic factors, etc.) and extrinsic characteristics (*e.g.*, environment, diet, stress level, etc.) can have widely divergent propensities to develop a particular disease condition or to discover why two similar individuals that are afflicted with the same disease condition may have their condition respond entirely differently to the same therapeutic treatment. There are likely a great number of factors that account for such differences. But even with our increased understanding of some of these factors, there is currently no quantitative tool that can predict how a patient (or a condition) will respond to treatment. To the extent that some diagnostic and prognostic tools are available, there are none robust enough to account for both the intrinsic and extrinsic characteristics

described above, the wide array of pathological conditions that may confront a patient, or the manner in which those conditions are likely to change with time and treatment.

One noteworthy exception is described in U.S. patent application serial No. 10/294,270, filed November 14, 2002. This reference describes systems and methods for diagnosing and treating various physiological conditions based on the comparative analysis of proteins found in, by way example, an individual's serum. This is accomplished by performing pattern recognition analysis and/or its subdisciplines (*e.g.*, discriminant analysis, feature extraction, error estimation, cluster anaylsis or "statistical" pattern recognition, and grammatical interference and parsing or "syntactical" pattern recognition) of the individual's protein profile with respect to, for instance, profiles from patients with a similar disease condition. This reference also describes the comparative analysis of other patient information to optimize a therapeutic treatment regimen. This is an application of proteomics -- the systematic analysis of the proteins expressed in a particular system at a given point in time to assess dynamic changes in the proteome, such as those associated with internal or external system perturbations. The human proteome includes all human proteins encoded by the genome (in all their modification states, *e.g.*, phosphorylated, etc.). The proteome continually changes in response to a vast array of factors (*e.g.*, internal/external events, disease, drugs, mood, etc.), while the genome remains relatively static and well-defined for an organism.

One significant obstacle in implementing the technology described in the aforementioned application as well as related technologies that involve the storage and comparative analysis of patient clinical information is the enormous processing and data storage requirements for such a system; particularly when such a system is implemented on a large scale. Such "large scale" implementation may involve, by way of example, a number of medical and other institutions, data obtained directly from multiple patient populations and/or obtained from many published clinical studies, and the routine use of the subject technology in doctor's offices (*e.g.*, via the Internet or other modes of electronic communication). These factors are just a few examples of how the subject technology may rapidly expand beyond the practical capacity of currently available computer processing systems and conventional configurations of database architecture. In short, the technology may surpass that which can be supported and efficiently handled by current computer systems.

Furthermore, currently available patient management and care systems use mostly unaltered data directly from patient tests to populate the underlying databases that house patient information. Thus, one of the shortcomings of conventional medical data management is the large number of ways that similar or identical concepts are described; for example, two

physicians may refer to a similar location of a tumor mass with differing descriptions (*e.g.*, "proximate to" or "nearby" a patient's liver, reference to a leg injury as "fractured" or "broken"). These variations in the data present a challenge to managing data sets in such a way as to have consistency among the data elements (*i.e.*, data descriptors or metadata). A lack of consistency can render it problematic or even impossible to use this data for any useful automated or computerized application. Currently, there is some use in the art of common data elements to solve this problem. These common data elements are used as metadata descriptors, and in effect create a common language to describe certain metadata. However, use of common data elements is limited in its breadth and implementation.

As medical technology progresses and increasingly relies upon computerized and other electronic systems to manage patient data, to analyze clinical information and to improve the accuracy of disease treatment, additional opportunities continually arise to better understand disease and to more thoroughly manage patient care. The diagnostic and therapeutic potential associated with an electronic representation of a patient's serum protein profile or the ability to offer more efficient patient management through the use of computerized analytical tools and database schemes remains largely untapped.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 illustrates a logical design of the database architecture of an embodiment of the present invention.

Figure 2 illustrates a network layout illustrative of an embodiment of the present invention.

Figure 3 illustrates a block diagram of an embodiment of the present invention, including a common data element maintenance and electronic data capture system.

Figure 4 illustrates a block diagram of the common data element repository according to an embodiment of the present invention.

Figure 5 illustrates a block diagram of a system including the web form engine, according to an embodiment of the present invention.

Figure 6 illustrates a process utilizing the web form engine to collect patient data via the Internet, according to an embodiment of the present invention.

Figure 7 illustrates a block diagram of a visit scheduler, according to an embodiment of the present invention.

Figure 8 illustrates a process which may be used to schedule patient visits in a clinical trial and to define new data elements, according to an embodiment of the present invention.

Figure 9 illustrates a fractionation of serum samples in connection with pattern recognition analysis, according to an embodiment of the present invention.

Figure 10 illustrates a variation in m/z read points across a spectrum among mass spectrometers, according to an embodiment of the invention.

Figure 11 illustrates a database architecture of a pattern recognition analysis technology, according to an embodiment of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

The present invention relates to a system and method for patient data and treatment management. It integrates a variety of scientific methods, devices and protocols. It also allows for a new means by which to approach medical treatment, drug discovery and the laboratory analysis of patient data. In general, the invention may include a computer system configured to store, manipulate and analyze medical data. As will be readily appreciated by those of skill in the art, the array of computer systems, research tools, databases and other sources of information that may be used in connection with the system and methods of the present invention to accomplish its objectives is vast. Nearly any hardware, firmware, software, operating system, database platform, networking technique or other conventional computer tool can be configured to operate in connection with the system and methods of the present invention, as will also be appreciated by those of skill in the art. Indeed, the scope of the type of devices that may be employed in connection with the system and methods is nearly boundless. Such devices may include those typically found in the traditional computer arts (*e.g.*, computer systems, networking tools, etc.), in medical diagnostics (*e.g.*, magnetic resonance imaging or "MRI," computerized axial tomography or "CAT scans," blood serum analyses, etc.), in therapeutic or surgical interventions, in research environments (*e.g.*, output of mass spectrometry, DNA sequencers, etc.), and beyond. The present invention is thus illustrated by the ensuing description of its components and particular features that may used in connection with the same, but it is in no way intended to be limited by the specific devices or systems described herein.

In the present invention, patient and/or clinical data may be maintained in one or more databases or repositories. These databases may be selected from any type of database presently known in the art or hereinafter developed. By way of example, the database or databases may be selected from extended markup language ("XML") databases, Oracle databases, Microsoft MySQL databases, or any other suitable databases. The databases may be constructed using a language such as XML or may alternatively be constructed by direct input of data. The use of object-oriented databases is also possible, and is considered as being within the ambit of the

present invention. One embodiment of a database architecture suitable for use in connection with the present invention is described below and is depicted in Figure 1.

Moreover, the databases and repositories of the instant invention may be stored in the same or different physical databases. By storing the databases and repositories in the same physical database(s), it may be possible to reduce administration costs; although there are benefits to separating the databases and repositories into different physical databases. Any database or repository may be split from a single physical database at any time.

The system of the present invention may be adapted for configuration on a network, such as an intranet (*e.g.*, a local area network). This may enable access to any or all of the databases of the invention, for example, within a hospital or system of hospitals. It may additionally or alternatively be adapted for configuration on the Internet (*e.g.*, to provide access to the database by remote users that may be unaffiliated with the owner of the database); remote terminals may thereby access the database. In various embodiments, remote users may subscribe or otherwise pay for such access to the database or to various features of the invention.

Metadata Repository

As illustrated in Figure 1, in one embodiment of the present invention, the system includes a metadata repository **101**. The metadata repository **101** may store a large quantity of metadata, which can be interpreted or analyzed by other components in the system. The metadata stored in the metadata repository **101** may include any desirable type of information about a patient, treatment, condition or any other information that may be useful in the course of patient management, diagnosis or treatment.

The metadata repository **101** can be based upon any type of repository known in the art. For example, but without limitation, it may be a repository based on the ISO/IEC 11179 repository standards. Furthermore, the metadata may be stored entirely in one common language or in a combination of languages, depending upon the particular configuration of the system. For example, for various medical applications, the metadata may be stored in the standard Health Level 7 ("HL7") language or any other language that may be convenient for describing the type of medical and clinical elements that may populate the metadata repository **101**. Use of a pre-existing language may allow users to store a large amount of metadata without having to define each data element. However, an entirely new or different language may be used to store metadata in alternate embodiments of the present invention.

In an embodiment of the present invention, new elements for which there is presently no corresponding defined element in the language being used in the metadata repository **101** may

be defined to work within the language. By adding new elements to an existing language, the capacity of the metadata repository **101** to store relevant information may be expanded to a broad array of further applications. As an example, in those instances where HL7 is the language used in the system, elements are already available for a large number of defined data types; for instance, data types commonly used in patient care, patient referral and hospital administration. If, however, an element is not available for a particular type of data (*e.g.*, the isotopic mass of a protein found in a patient's blood serum), it is within the scope of the present invention to create a new element for such data that operates with the HL7 language. In this manner, the present invention is not limited by any particular type of information that can be stored in its metadata repository **101**.

More particularly, metadata in the metadata repository **101** may include any type of data element. For example, it can include existing data elements from other repositories, such as the National Cancer Institute Center for Bioinformatics repository. The metadata repository **101** also allows storage of basic patient and clinical data, such as, by way of example, age, sex, blood type, weight, prostate specific antigen ("PSA") score, stage of cancer, type of treatment regimen, clinical findings from an office visit or any other type of patient or clinical data. Other types of information, such as the profile of serum proteins found in a patient's serum sample at a particular time and/or the results of the comparative analysis of that profile with others in the course of an analytical process may also be included in the metadata repository **101**.

Data stored in the metadata repository **101** may be organized by patient, date of office visit, or in any other manner that may be suitable to accomplish the particular goals of an embodiment of the inventive system. The data may include results of clinical trials or other broad based tests. They may also include data collected from a hospital electronic records system as well as data obtained from published clinical trials or nearly any other source. There is essentially no limit to the types of data that may be included in the metadata repository **101**. The storage capacity of the metadata repository **101** may be increased or supplemented to account for growth.

The metadata repository **101** may further be linked to or integrated with any medical or research instrument such that the output from such an instrument is input directly (*i.e.*, via instantaneous electronic communication) or indirectly (*e.g.*, by being downloaded or otherwise obtained from the instrument or its associated storage media to the metadata repository after the instrument is used in connection with a medical or research procedure) into the metadata repository **101**. By way of example, medical instruments that may be particularly suitable for use in connection with the system and methods of the present invention may include, but are in

no way limited to, mass spectrometers, gas chromatographs, capillary electrophoresis instruments, MRI or CT imaging equipment, DNA sequencers, microarray gene expression tools, etc.

In one embodiment, raw data involving a particular patient is collected during a clinical trial, in the course of a routine physical examination with a treating physician, or at any other time or in any other manner, and thereafter entered or otherwise input (*e.g.*, automatically, when the data is generated from a research instrument) into the metadata repository **101**. The data may be converted into an XML format. Once this raw patient data has been input, it may be stored in a central database.

Further to the above discussion of sources of information and devices that may be used in connection with the system and methods of the present invention, there are still further methods for data input, manipulation and initial editing that are within the invention's purview. For example, an editing tool may be used to edit the repository of metadata. The editing tool may be a metadata curation tool that is used to browse and maintain the metadata. The metadata curation tool may, for example, be used to edit or delete incomplete metadata entries. Alternatively, the editing tool may be a form authoring tool that browses data elements stored in the repository and allows authors to build forms from common data elements. The forms can then be rendered on paper using a render agent, such as the INFORM Electronic Data Capture System (available from Phase Forward, Inc.; Waltham, MA) or any suitable web form engine. Furthermore, a form authoring tool may be used for quality control purposes. Such a tool can validate data or help define queries to obtain an accurate database. Data may also be input via input **110** by text and/or language processing. For example, data may be input directly into the metadata repository **101** as text, or a processor may first convert text or language into data elements. Any electronic medical records interface may be used to input data in connection with an alternate embodiment of the present invention.

In one embodiment of the present invention, a web form engine that reads form definitions generated by the form authoring tool is used to drive a web application for electronic data collection. This embodiment is illustrated in greater detail in the Examples section below, along with further embodiments of the invention that may be employed in creating and using forms.

Facts Database

In another embodiment, as depicted in Figure 1, the present invention may include a facts database **102**. The facts database **102** may be a separate, standalone database or may be

integrated with any of the other databases of the invention, as will be readily appreciated by those of skill in the art. The particular manner in which the facts database **102** is configured with other system databases as well as other system components may be selected based on a variety of factors, such as, by way of example, system efficiency, electronic storage capacity, database security, cost, ease of use and the like. In general, the facts database **102** is included in the system to promote its operational efficiency (*e.g.*, by aiding in the standardization of raw data contained in the metadata repository **101** or by imparting an organizational scheme to this raw data that renders it more easily accessible by other system components and/or software programs). Operational efficiency may significantly facilitate the various analytical processes and machine learning protocols that may be performed with respect to the information stored throughout the system (*i.e.*, in the metadata repository, the database of facts, etc.). This may be particularly important in the context of the instant invention when its teachings are applied to the medical field, because, as previously noted, the system is likely to contain a tremendous quantity of information and system efficiency may be at least partially compromised without procedures and/or components in place to address this issue. Indeed, there are many means for doing so, but the facts database **102** is a helpful tool in this regard.

More specifically, the facts included in the facts database **102** may be derived from the raw data stored in the metadata repository **101**. As described above with respect to the type of information that may be housed in the metadata repository **101**, correspondingly, the facts included in the facts database **102** may similarly represent a host of different types of information about a patient (or a population of patients). Generally speaking, the most significant difference between the raw data in the metadata repository **101** and the facts included in the facts database **102** is that the facts may be more readily interpreted, compared and analyzed by various features of the inventive system or by the many external software applications and analytical programs that may be used in connection with it. As will be readily appreciated by those of skill in the art, the array of modifications and/or reorganizations that may be implemented in translating the raw data in the metadata repository **101** into facts in the facts database **102** is vast. As such, any suitable translation, modification, reorganization, filtering of information or other operation performed with respect to the raw data in creating, updating or otherwise importing information into the facts database **102** may be implemented -- either alone or simultaneously, in series or in parallel, as a one-time or recurring event and/or in any desirable combination -- in connection with alternate embodiments of the present invention.

For example, there may be a need to initially standardize raw data that is collected from various published clinical studies, data that is generated by research tools (*e.g.*, with varying

sensitivities or units of measurement) or any other form of raw data that might benefit from some initial form of standardization prior to more substantive system analysis. As will be readily understood by those of skill in the art, the various manners in which the raw data in the metadata repository **101** may be standardized or otherwise manipulated in this context in importing it to the facts database **102** is practically unlimited. Therefore, any initial standardization or similar procedure that may be useful in connection with various embodiments is contemplated as being within the scope of the present invention, and can be readily implemented by those of skill in the art without undue experimentation.

In another example, there may be a need to filter raw data that is stored in the metadata repository **101**; for instance, when it is continually collected from a research tool that is integrated with the system. In one embodiment of the present invention, the system is used in connection with equipment that obtains information regarding the characteristics of proteins in an individual's serum sample. This is illustrated below in the Examples section. In this feature of the invention, the equipment may be integrated with the inventive system such that the raw data collected in the course of examining proteins is continuously uploaded to the metadata repository **101**. There may be particular items of information contained in that raw data that are relevant to accomplish the specific goals of the embodiment of the system (*e.g.*, to study a particular disease that is correlated with the presence of a finite number of proteins found in the serum). As such, only a portion of the raw data may be needed for further analysis by the system, and the facts database **102** may thus be configured to include only that information required for further analysis. The remainder of the information may remain stored in the metadata repository **101** as raw data.

In another example, and in accordance with a still further embodiment of the present invention, raw data contained in the metadata repository **101** may be "reorganized" and imported to the facts database **102**; independently of whether any other operation (*e.g.*, a standardization operation) is performed with respect to this raw data. For instance, among the raw data contained in the metadata repository **101** may be patient information obtained during a patient's office visit (*e.g.*, body temperature, blood pressure, weight, etc.). Notwithstanding any difference in customary units of measure employed in various parts of the world (which can be readily accounted for in a standardization operation), in the United States at least, body temperature is generally measured in degrees Fahrenheit ("°F"), blood pressure is measured in millimeters of mercury ("mmHg") (*i.e.*, with respect to both systolic and diastolic arterial blood pressure) and weight is measured in pounds ("lbs"). The reorganization of raw data from the metadata repository **101** may simply involve importing this information into the facts database

102 such that the raw data is associated with another data field, such as a patient's name or a alphanumeric identifier assigned to the patient in the course of a clinical study (*e.g.*, to preserve patient confidentiality). While this information may be somehow associated in the metadata repository **101**, it may be streamlined in the facts database **102** by separating it from other information also associated with this patient that may not be of particular use for the analytical processes used by the system in a particular embodiment. The information may be re-associated at a later point in time by obtaining the same from the metadata repository **101**, if that becomes desirable. However, this may significantly reduce the volume of information in the facts database **102** as compared with the metadata repository **101**; thereby improving system efficiency when an analytical program queries the facts database **102**. As will be readily understood by those of skill in the art, the various manners in which the raw data in the metadata repository **101** may be reorganized, streamlined or otherwise manipulated in importing it to the facts database **102** is practically unlimited. Therefore, any reorganization that may be useful in connection with various embodiments is contemplated as being within the scope of the present invention, and can be readily implemented by those of skill in the art without undue experimentation.

As noted above, a host of data operations may be implemented in translating information from the metadata repository **101** to the facts database **102**. In still further embodiments of the present invention, information may be imported directly into the facts database **102** without first being input to the metadata repository **101**, or information may be imported identically and directly to both the metadata repository **101** and the facts database **102**. The most convenient and/or appropriate means for importing information into the metadata repository **101** and/or the facts database **102** may be selected and implemented by one of skill in the art to accomplish the goals of a particular embodiment of the present invention without undue experimentation.

Moreover, the system and methods of the invention may be fully operational without the inclusion of a facts database; however, a facts database may render it possible to store patient data in a database that is oftentimes smaller than the metadata repository **101** itself. This may provide easier and more efficient access to raw data and to the true substance of information housed in the system. In alternate embodiments of the instant invention, information (*e.g.*, patient information) can be obtained and analyzed: (1) directly as raw data from the metadata repository **101**, (2) directly as facts from the facts database **102**, or (3) in any combination as raw data and facts from the metadata repository **101** and the facts database **102**, respectively.

Rules Database

In another embodiment, and as illustrated in Figure 1, the present invention may further include a rules database **103**. Generally speaking, the rules database **103** may "interpret" information contained in the facts database **102** and/or contained in the metadata repository **101**. Rules included in the rules database **103** may be selected from a range of different types. For example, in the medical context, the rules can include generally accepted principles regarding diagnosis and treatment (*e.g.*, that a PSA score within a particular numeric range correlates with a certain likelihood of a patient having prostate cancer). In another embodiment, the rules in the rules database **103** may also account for variation among the raw data contained in the metadata repository **101** and/or the facts contained in the facts database **102** by performing a substantive standardization operation.

For example, in one embodiment, the metadata repository **101** and/or the facts database **102** may include information input by a physician regarding his diagnostic description of a patient's tumor mass. Specifically, one physician may input that his patient's tumor is located "nearby" the prostate. Another physician diagnosing a similar condition in a different patient may input that her patient's tumor is located "next to" the prostate. A third physician diagnosing yet another similar patient may input that his patient's tumor is located "adjacent to" the prostate. As will be readily appreciated by one of skill in the art of oncology, each of the three aforementioned patients has a similar condition: a tumor mass in close proximity to the prostate. However, a computerized analysis of this information may not be savvy to the alternate ways in which this identical condition might be described by medical practitioners. Moreover, if an analytical program were taxed with the chore of first identifying and accounting for these divergent descriptions prior to any further analysis, the efficiency of that analytical program may be compromised. Therefore, in this embodiment, it is an object of the rules database **103** to account for the substantive variation among diagnostic descriptions in the information contained in the metadata repository **101** and/or the facts database **102**.

Using the example of the three patients and three physicians from above (each of whom has diagnosed the same condition, but has described it in a different manner), a rule in the rules database **103** may recognize the differing descriptions of this diagnosis and "translate" the information (*i.e.*, that one patient has a tumor located "nearby" his prostate, a second patient has a tumor located "next to" his prostate and a third patient has a tumor located "adjacent to" his prostate) accordingly. For example, for each of these patients, the rules database **103** may indicate that each of the three patients has a tumor located "adjacent to" his prostate. With this identical description applied to each of the three patients, it may be markedly more efficient to

now implement an analytical tool that can study the information about these patients, because the discrepancy resulting from the various physicians' particular descriptions of the identical physiological condition has been substantively standardized.

The type of "initial" standardization operation described above with respect to the facts database **102** differs from the type of "substantive" standardization operation that may be implemented by a rule in the rules database **103**. Although there is not necessarily a bright line distinction between these types of operations, generally speaking, an "initial" standardization operation that may be implemented in the facts database **102** might account for mathematical or numeric variation among the raw data; for example, variation in units of measurement, variation in the sensitivity of research tools or diagnostic equipment, variation in the descriptor used to describe a patient's sex (*e.g.*, "M" or "male" being the same) or the like. On the other hand, a "substantive" standardization operation performed by a rule in the rules database **103** may address variation in patient information that is not so easily "converted" (*i.e.*, a more critical analysis is required for the information to be standardized).

The rules database **103** may also include "soft rules." Soft rules may vary subject to an ongoing analysis of the information (*e.g.*, raw data, facts) included in the system, such as by the implementation of "machine learning," described in greater detail below. This is most likely to take place over time, as additional information is input to the system and continuing analysis is performed. Soft rules may additionally or alternatively include rules based upon the results of clinical studies or those gleaned from accounts in scholarly publications. The soft rules may also include information derived from research conducted within a system user's organization (*e.g.*, a research hospital or pharmaceutical company), and/or they may include information from external sources. Soft rules may be included in the rules database **103** or may be housed in a separate soft rules database (not shown).

As will be readily understood by those of skill in the art, "machine learning" generally refers to a broad class of probabilistic and statistical methods for estimating dependencies between data and using the estimated dependencies to make predictions. It can be accomplished by simple algorithms that look at a two-dimensional distribution of data points and create a rule to define the distribution. It is also possible to begin with a rule, graph data points output from the rule and then look for clusters of data. If there is one cluster of data points that exists at a large margin away from other data points, the rule is likely to be a good rule (*i.e.*, one that accurately describes a feature of the data). If there is a small margin, the rule is less likely to be a good rule. The best rules generally depend upon only a few critical sample points, referred to as "support vectors." Machine learning machines relying on such points are commonly referred

to as support vector machines; they may twist and stretch the space before separating data points to further improve analysis of rules.

More complex machine learning algorithms may involve plotting three-dimensional distributions of data points. For example, in principal component analysis ("PCA"), the machine finds the directions that explain most variation in the data. It projects all of the data onto these "important" directions. Then, this smaller projection of data is used to run the simple vector machine. It is also possible to twist and stretch the data before PCA. Any form of machine learning algorithm (*e.g.*, kernal PCA, including, for example, polynomial kernels, Gaussian or radial basis function, sigmoid, neural networks) may be used in connection with various embodiments of the instant invention, as will be readily appreciated and implemented by those of skill in the art without undue experimentation.

In another embodiment of the present invention, users of the inventive system may update or register their own soft rules to create a personalized system. For example, a particular physician may treat her patients in accordance with diagnostic criteria that are not otherwise accounted for in a particular embodiment of the system. For instance, the system may provide certain information based on specifically accepted correlations between ranges of PSA score and likelihood of having prostate cancer, as described above, while the particular system user may have a more restrictive view than that which is generally accepted based on her experience or opinion. As such, a soft rule may be used to account for the physician's personal diagnostic criteria. Furthermore, in various embodiments of the present invention, the soft rules that are implemented may be limited to those selected by a system user. For example, the user may choose to apply only her soft rules, all available soft rules or any combination or permutation thereof. Additionally, the system may allow individuals other than the system user(s) to register rules. This feature may also be dependent on the preference of the user.

The rules database **103** may be changed and/or updated over time to account for new developments in knowledge of disease and treatment and the like. These changes may be input via input **110** manually, or may be derived from machine learning built into the system. The rules, including the soft rules, can be associated with the system's machine learning capabilities and processes.

Output from the rules database **103** can be used for any number of purposes, as will be readily appreciated by those of skill in the art and implemented without undue experimentation. As such, the system output from the rules database **103** may be exported to a host of different components or systems, whether these components or systems are directly integrated with the inventive system (*e.g.*, with an integrated software suite, through a network connection, via the

Internet) or exist separate and apart from the system (*e.g.*, requiring an export of information from the system and simultaneous or subsequent import/upload to the separate system). For instance, the output may be sent to a forms engine to prepare various forms, as is routine practice in the medical field. The output may be used for internal machine learning (*i.e.*, by the system) to assist in generating new rules and/or in refining those already present in the system (*e.g.*, in defining new standards of patient care or updating those that are already in use). The output may also be used for scheduling appointments and/or treatment regimens for patients. For example, a rule may provide that a cancer patient undergoing a particular course of treatment should receive the subject treatment once every two weeks for three months (*e.g.*, chemotherapy), and/or it may require weekly office visits throughout the duration of the treatment. The output may also be utilized by rules in the processes database **104**, described below.

Furthermore, the system may indicate a "preference" for a particular diagnostic or treatment option based on the machine learning and other rules and analyses used in accordance with various embodiments of the present invention in the rules database **103**. By way of example, as described in U.S. patent application serial No. 10/294,270, the system may implement a pattern recognition technique with respect to a pattern of proteins in a patient's blood serum, as compared with the protein patterns identified in other patients' blood sera. This may be particularly advantageous in instances where a number of treatment options are available for a particular patient (*e.g.*, the variety of pharmaceutical "cocktails" conventionally prescribed for the treatment of HIV/AIDS). If the protein pattern in the patient's blood serum is more closely correlated with patients that successfully responded to one standard drug regimen as opposed to those that successfully responded to an alternate drug regimen, then the system may suggest that the patient be treated with the former rather than the latter. Other applications for this pattern recognition technique are more thoroughly described in U.S. patent application serial No. 10/294,270, and the use thereof in the context of the present invention will be appreciated and may be readily implemented by those of skill in the art without undue experimentation.

The rules database **103** is a flexible and dynamic tool, which can be updated, altered and/or supplemented as desirable to accomplish a broad array of tasks. It will be readily appreciated by those of skill in the art how one might further utilize this feature of the invention to facilitate the treatment and/or diagnosis of illness, or more generally to improve patient care and management.

Processes Database

In a still further embodiment, as illustrated in Figure 1, the present invention may include a processes database **104** that contains an array of processes. In various embodiments of the invention, processes may describe particular features derived from standards of patient care (*i.e.*, widely accepted treatment protocols that are commonly used by physicians in a particular field in treating/diagnosing illness in their patients). For example, specific processes may indicate that a particular test should be run on a patient, that a certain medication should or should not be prescribed to a patient, or that a medication should be prescribed in a particular quantity/dosage/frequency. The pattern recognition methodologies described in U.S. patent application serial No. 10/294,270 may also be particularly useful in this regard, and may be readily implemented in connection with the processes database **104** of the instant invention.

The selection of a particular process (or processes) may be dependent upon the outcome of an analysis performed by the system based upon a rule (or rules) that examines various facts and/or raw data. For example, the raw data in the metadata repository **101** may indicate that a patient is male, 64 years of age and has a PSA level of 19 ng/mL. A series of facts in the facts database **102** may indicate that the patient has a tumor located "adjacent to" his prostate but no other detectable tumor masses in his body (*e.g.*, based upon the results of a CAT scan, MRI and/or ultrasound). A rule in the rules database **103** may, upon analysis of these raw data and facts, indicate that the patient has a high likelihood of having early stage prostate cancer. Based on the raw data in the metadata repository **101**, the facts in the facts database **102** and the output of the rules in the rules database **103**, a process in the processes database **104** may indicate that the patient should be started on a particular treatment regimen (*e.g.*, external radiation therapy, five times per week for four weeks).

As with other database components of the present invention, the processes database **104** may be a separate, standalone database or may be integrated with one or more of the other databases used in connection with various embodiments of the present invention based on the same criteria set forth with respect to the other system databases.

EXAMPLES

The following examples describe a range of applications of the system and methods of the present invention, as well as a number of components that may be readily integrated and/or otherwise used in connection with the same. These Examples demonstrate some of the many configurations of the system of the invention, and the potential impact it may have on the conventional practice of medicine.

EXAMPLE 1

Implementation of Pattern Recognition Analysis

The present invention is integrated with diagnostic equipment that enables the implementation of pattern recognition analysis of a protein profile, such as a serum protein profile, as described in U.S. patent application serial No. 10/294,270. A wide array of samples may be obtained and used in conjunction with alternate embodiments of the system (*e.g.*, a body fluid, such as blood, plasma, serum, CSF (spinal fluid), urine, sweat, saliva, tears, breast aspirate, prostate fluid, seminal fluid, stool, cervical scraping, cytes, amniotic fluid, intraocular fluid, mucous, moisture in breath, animal tissue, cell lysates, tumor tissue, hair, skin, buccal scrapings, nails, bone marrow, cartilage, prions, bone powder, ear wax, etc.). In obtaining samples for analysis with the system, a patient may be subjected to a stress test prior to obtaining a sample (*e.g.*, inhale a breath of smoke and obtain a sample from the exhalate, feed a patient a particular diet prior to sample harvest), or samples may be obtained from the patient both before and after a stress test, such that the effect of the stress on the patient may be assessed. Moreover, a sample may be perturbed in some manner after being extracted from a patient (*e.g.*, heating a sample, exposing a sample to UV light, adjusting the pH of a sample, allowing time to elapse with the sample *ex vivo* before analysis).

A standard control serum for particular applications of the system is useful in data analysis, although not required (*e.g.*, a control serum may contain trypsin, a protease, a glycosilase or another enzymatic manipulant to cleave proteins in a predictable manner before further analysis). Alternatively or in addition, a laser (*e.g.*, a 25W laser) may be used to "break up" proteins included in the sample.

Once a sample is obtained, it is prepared for further analysis. In one preparation protocol, the proteins in the sample are chemically reduced (*e.g.*, with dithiothreitol, or "dTT"), denatured (*e.g.*, with guanadine HCl pH 6, or urea), alkylated (*e.g.*, with iodoacetamide, or "IAA") and cleaved (*e.g.*, with trypsin). Once digested, an isotopic labeling scheme may be applied; for example, a control sample may be labeled with a hydrogenated molecule, while a test sample may be labeled with an isotopic deuterated molecule. These isotopically labeled samples may then be mixed prior to analysis. If no control is included, labeling may not be necessary; although resultant intensities may be less reliable without a control for relative quantitation or comparison.

After sample preparation, the system includes a liquid chromatography ("LC") phase followed by mass spectrometry (*i.e.*, "LC/MS"). The mass spectrometry phase is performed using a Fourier Transform Ion Cyclotron Resonance Mass Spectrometer ("FTICR-MS"). The

LC phase can be forward or reversed phase, and may be of one or several dimensions (*e.g.*, 1D, 2D, 3D, etc.) including size exclusion and ion exchange chromatographic techniques. It may also be standard LC, high performance LC ("HPLC") or fast performance LC ("FPLC"); depending upon system parameters, the character of the sample and the desired format of system output. Reverse phase HPLC is employed in this instance.

Further, electrospray ionization ("ESI") is used online with an HPLC in the system. To obtain better system resolution, after passing through a standard ZSPRAY ionization source (available from Micromass Ltd.; Manchester, UK), particles pass through a triple quadrupole followed by a hexapole prior to entering the capillary of an FTICR. Microfluidics may also be used after the LC phase and before MS to facilitate system scale-up, as a greater number of analyses can be performed simultaneously.

In order to overcome the system bottleneck created by running samples through a single magnet, in series, in a modified version of the system, a bi-directional LC/MS apparatus is employed. Two ICR cells are inserted into a single magnet (*i.e.*, each configured 180° from the other, facing opposing ends of a cylindrical magnet), and samples are introduced independently, with respect to each cell. This doubles system efficiency, as data from two samples may be obtained in one magnet with two detectors. This may be particularly advantageous in terms of system scale-up. Alternatively, a "machine gun" ionization spray may be used to rapidly fire different sprays into a single magnet. In yet another alternative design, portions of a sample are fired into regions of a magnet other than or in addition to the magnet's absolute center. In this manner, many portions of a sample may be analyzed simultaneously. For example, the molecular dynamic range or m/z ratio range may be divided into arbitrary units, and each unit range can be fired into a different region of the magnet (*e.g.*, Range #1 fired at Region #1, Range #2 fired at Region #2, etc.). While the raw data obtained from such a procedure is distorted, the distortion is predictable, because each sample portion of a particular range is always fired into the same region in the magnet. Thus, the distortion is accounted for with an appropriate mathematical correction. By way of example, wherein orbits are perfectly elliptical, the detected cyclotron signal data is transformed into mass spectra by applying elliptical functions rather than the spherical functions of the basic forward Fourier transform.

Between the LC and FTMS phases, a fractionation step may be integrated into the system (referred to as "LC/MS Plus" or "LC/MS$^+$"). In this version, and as illustratively depicted in Figure 9, a flow splitter **903** is included between the LC **902** and FTMS **904** phases to collect fractions by a fraction collector **905** from the sample flow out of the LC. Because the fractions obtained are correlated with retention time (*e.g.*, a mass spectrum of interest produced

at 34 minutes correlates with the fraction obtained at 34 minutes), a fraction of particular relevance may be readily identified. This fraction may thereafter be subjected to further analysis. For example, the proteins in a fraction of relevance may be identified by protein sequencing; a factor (*e.g.*, a metal-specific tag such as silver atoms) may be bound to the sugars in the fraction of relevance as a means of examining post-translational modifications; protein-protein interactions may be examined; or, Raman spectroscopy may be employed to indicate any chemistry in the fraction, such as the presence of silicone or gold molecules.

An analysis of the fractionated sample may be automated, such that it occurs in parallel with the FTMS analysis. Thus, a peak of interest that is identified by pattern recognition may be immediately coupled with additional information about the specific proteins and other molecules associated with that peak. A vial including the fractionated sample may then be used in other applications. For instance, the vial containing that sample may be sent to an outside research entity for use in the development of diagnostic or therapeutic methods (*e.g.*, a certain number of factors are identified as important for ovarian cancer, and vials of those factors obtained from the flow splitter are provided to a pharmaceutical company for drug testing and development). Rather than sending the physical vial, information obtained by subjecting the fractionated sample to additional testing may be coupled with the mass spectrometry data and stored in the system database, or it may be transmitted to an interested party via, *e.g.*, the Internet.

By automating the fractionation step, additional features may be included in the system. For example, after generating FTMS data on the entire sample, the system may perform further analyses on those proteins that it has identified as, *e.g.*, the five most abundant in the sample. Another example includes the system automatically picking the peptides of particular masses when a certain spectrum is obtained (*e.g.*, when spectrum 122 is obtained, the peptides with mass 100, 250 and 322 are analyzed).

FTMS is a primary component of this separation system; however, it can be used in combination with an array of other techniques and systems to identify peaks of relevance for analysis. As previously noted, HPLC may be used in the system. In alternate embodiments, the system may employ a double iteration of mass spectrometry (*i.e.*, MS/MS); it may combine mass spectrometry with Raman spectroscopy (*i.e.*, MS-Raman); it may incorporate both liquid chromatography and Raman spectroscopy with mass spectrometry (*i.e.*, LC-Raman-MS); or it may include a double iteration of liquid chromatography with mass spectrometry (*i.e.*, LC/LC/MS). In fact, any combination of tandem LC and MC may be utilized. Other variants of this system may be used, depending on the particular application, the form of the sample being analyzed and the particular aspects of data analysis that are desired. For example, alternatives to

FTMS may include Matrix-Assisted Laser Desorption Ionization with Time of Flight ("MALDI-TOF"), the API 3000 LC/MS/MS System (available from Applied Biosystems; Foster City, CA), the QSTAR XL Hybrid LC/MS/MS System (also available from Applied Biosystems) or Q-TOF (available from Micromass).

As an alternative to LC/MS, a modified Surface-Enhanced Laser Desorption Ionization ("SELDI") technique may be implemented. SELDI is a chip-based molecular imaging process, in which antibodies are dispersed (*i.e.*, cross-linked) on the surface of stainless steel plates provided by the manufacturer (available from Ciphergen Biosystems, Inc.; Fremont, CA). The stainless steel plates may be modified such that they do not include antibodies on their surface. This modification provides superior resolution beyond that which can be achieved with conventional SELDI plates. It may be particularly desirable to use stainless steel plates with some molecule or combination of various molecules dispersed thereupon other than the antibodies used in connection with the commercially available SELDI plates.

In general, pattern recognition algorithms can process data that are input in vector form (*i.e.*, a list of numbers that is always the same length). Standard LC/MS readout lends itself to vector representation suitable for analysis with a pattern recognition algorithm. However, the LC readout itself can be used, as can any additional data that is desirable for a particular embodiment of the program. Clinical outcome data and other information can be attached to the vector. A standardized annotation language for this information (*e.g.*, HL7) facilitates database analysis.

Differing resolutions among readouts obtained from LC/MS (*i.e.*, different numbers of frames from either the same or different equipment) can produce "lists" of numbers that are not the same length. However, this does not present an obstacle to vector representation and pattern recognition, because m/z values can be mapped frame-to-frame to account for any resolution-based discrepancy. Furthermore, as illustratively depicted in Figure 10, actual m/z read points across a spectrum may differ among mass spectrometers. To obtain data at a set of standard read points, and thus enable optimal comparison of patterns, data that is read from non-standard read points can be interpolated to standard read points (*e.g.*, by "initial" standardization of raw data in the system). This may be particularly useful when data sets from a number of facilities using different equipment are input to the system. Linear interpolation may be sufficient based on the jagged data profiles produced by LC/MS of serum proteins, although more complex forms of interpolation (*e.g.*, smooth interpolation, quadratic interpolation, spline interpolation, wavelets) may be effective in other instances. Additionally, the calibration file from remote equipment should be provided to aid in importing data to the system.

As data is collected in the FTMS phase, a Fourier transform is applied and the resultant data is thereafter exported through a digitizer in binary form to the data analysis component (*i.e.*, computer) of the system. In a modified version of the system, the data is exported in ASCII form, rather than binary. Data exported in binary form is smaller in size than ASCII data, and is already in the format that conventional analysis programs generally require. However, to the extent that the various computer applications may require ASCII input, the system is versatile enough to meet that need, as well.

The system may be further streamlined by automating the export of data to the computer, and by using a faster digitizer (*e.g.*, a 14 bit digitizer may be sufficient for particular applications, but a 32 bit, 64 bit or faster digitizer may be advantageous in alternate embodiments). Moreover, the computer may be replaced with a component that acts as a client to the system database. By doing so, samples are processed with FTMS and the resulting data is automatically transferred to the data analysis component of the system, where it is analyzed and stored, as discussed in greater detail, below. A printer may additionally be included in the system, such that a hard copy of the data produced by FTMS may be generated prior to analysis and storage in subsequent phases of the system. This allows for validation of the FTMS phase.

As illustrated in Figure 11, raw data **1101** is input to the system, at which point a form of linear time fast filtering **1102** is utilized to provide a more useful view of the data. Additional operations may be included, such as peak extraction, peptide assignment and random projections (*i.e.*, a method of significantly compressing vectors in a way that preserves the output of machine-learning algorithms used in the system) (not shown). These operations provide different views of the vectors, and their output is channeled to a database that may be public, semi-public or private (*e.g.*, the facts database **102** of the system). Virtual samples may also be created at this stage to address issues such as calibration error, intensity error variation, scaling error and the like. This may decrease the reliance on specific scaling. The raw data may be stored in a database **1103** (*e.g.*, the metadata repository **101** of the system), as well.

Either in addition to or as a substitute for pattern recognition, image or object registration (*i.e.*, a form of image recognition) may be used in analyzing profiles obtained from mass spectrometry. Deconvolution is a conventional method used in many commercially available tools for performing this type of analysis, but by instead using a series of image recognition algorithms, deconvolution can be avoided.

Raw data input to the system can be stored in a CD, DVD or tape library. Disks are not likely to be an efficient storage medium for a system with roughly 20 GB of data for every sample. Moreover, algorithms that analyze data "closeness" (*e.g.*, a tape placement algorithm)

may be implemented to direct raw data to an optimal storage location, such that data values that are correlated with one another appear in the same data block. Automated storage of data based on domain knowledge may significantly improve data accessibility and system efficiency. An error correcting code may also be used to address reliability concerns. A backup system is also included to store a second copy of the raw data.

After initial processing and backup, machine-learning algorithms are implemented to identify groups of features in the data that move together, and which explain variations in the data. To the extent that these features have clinical relevance, they may be used as rules in the rules database **103** of the system. Support vector machines may accomplish the task of identifying groups of features in the data that move together by applying methods such as kernel PCA **1107**. Support vectors can make solid statistical statements about any incoming vector; they are not limited to proteins or any other specific molecule. To run support vectors, a centralized covariance matrix **1104** is included in the system to maintain information about the closeness of the relationship among the samples in the database. The database architecture thus includes a centralized covariance matrix **1104** with a suite of kernel PCA's **1106**, **1107** running on top of it, as well as a covariance database **1105**. Multiple covariance matrices may be included in alternate versions of the system (*e.g.*, covariance matrices in logarithmic form, semi-log form, filtered, non-filtered).

Once the covariance matrix is operational (notwithstanding the fact that it may be periodically updated as new data is generated), the important features of the data (*e.g.*, the rules) have been identified. As new data is input, the features (*e.g.*, the rules) are extracted into either the same or another database **1108** (*e.g.*, the rules database **103** of the system), where a feature view of the data is presented. Thus, there is a raw view of the data **1103** (*i.e.*, the metadata repository **101**) that is immediately processed, and a feature view of the data **1108** (*i.e.*, containing the biologically relevant information) (*e.g.*, in the rules database **103** of the system) is obtained. Supplemental clinical data and other patient information may be included in the database **1103** that includes the raw data.

The machine-learning algorithms primarily operate upon the feature view (*e.g.*, in the rules database **103**), although they may have access to the raw view (*i.e.*, in the metadata repository **101**) or other views (*e.g.*, in the facts database **102**), as well. The data may be reprocessed periodically to extract new features (*e.g.*, rules) and produce a new feature (or rule). Old features (*e.g.*, rules) may be maintained such that programs running on them are not disturbed, and to support any studies that have identified good information therein. This reprocessing can take place periodically as a static batch feature identification and extraction

that essentially creates a new version of the feature. Alternatively, the reprocessing can be dynamic.

The system can be readily linked to external databases including pertinent information (*e.g.*, Genbank, other databases containing biologic information). In addition, the databases that are made public or semi-public (*e.g.*, to query for research purposes, for publication) may be maintained on a standard platform.

As described above, there are a wide variety of applications for this technology. For example, it may be used to predict the development of a particular disease condition (*e.g.*, chance of having a heart attack, developing Alzheimer's Disease or cancer); to predict a response to a system perturbation (*e.g.*, how an individual will respond to smoking); to identify therapeutic targets for treatment of disease (*e.g.*, identifying protein peaks of relevance and directing treatment at the same); to predict whether an individual will respond to a particular therapeutic intervention (*e.g.*, in selecting participants for a clinical study); to identify the signature of an outcome and prepare a treatment, accordingly (*e.g.*, a side-effect in a clinical study); to identify individuals for whom a particular therapeutic agent might be toxic (*e.g.*, a diabetes drug which causes liver failure in a small percentage of patients); to select the most appropriate style of implantable device (*e.g.*, configuration of a particular artificial hip or knee); to detect bacteria (*e.g.*, detection of *H. Pylori* based on products found in breath moisture); to determine enzyme activity (*e.g.*, by comparing with a signature associated with high membrane metalloproteinase or human kallikrein-2 activity); to test for the presence of a gene mutation (*e.g.*, presence of BRCA1 mutation to predict breast cancer); to determine physiologic differences among races and other groups (*e.g.*, different reaction of Japanese individuals to alcohol as compared to Caucasian individuals); to determine if an individual has an allergy (*i.e.*, without having to expose the individual to the allergen); in prenatal testing (*e.g.*, to screen for genetic characteristics); in intelligence testing (*e.g.*, if particular proteins produced by the brain are associated with IQ); and to optimize dosing with a particular drug. Another application involves administering a drug to an animal and thereafter defining the proteomic pattern in the animal. One may then search for the peaks identified in that analysis in a human patient, and high throughput analysis can be implemented to determine whether the target proteins are present in the human. This may simplify the identification of a therapeutic target.

EXAMPLE 2

Common Data Element Maintenance and Electronic Data Capture

The system of the invention may include a common data element maintenance and electronic data capture feature. The common data element maintenance and electronic data capture feature provides a mechanism for simple, convenient maintenance and exploitation of a robust dataset, such as, for example, metadata. This may be particularly advantageous for research institutions, including hospitals, clinics, universities and various others that maintain large data repositories.

Figure 3 illustrates an embodiment of the common data element maintenance and electronic data capture system **300** according to an embodiment of the present invention. As shown therein, the common data element maintenance and electronic data capture system **300** includes a form authoring tool **304**, which may be used to create new forms, a render agent **308** to render the forms in various formats, processors **312** and **314** to execute the instructions of the form authoring tool **304** and the render agent **308**, a processes database **104** to store the forms created, and a common data element repository ("CDER") **310** to store the data.

The CDER **310** may be used to collect and store data. The CDER **310** may provide a set of standard questions and data representations that can be used to build case report forms and can be expanded to incorporate new data elements as necessary. This expandability feature allows for the definition of more standard data elements and avoids having data elements in the system with overlapping meaning.

Figure 4 illustrates a block diagram of the CDER **310**, according to an embodiment of the present invention. As shown therein, the CDER **310** may further comprise a metadata repository **101** to store the data and various data protection mechanisms.

Various data protection mechanisms may also be used within the CDER **310** to safeguard patient privacy by protecting the patient data, protecting the integrity of the data and providing an expedient means for data recovery in the event of a disaster. As shown in Figure 4, the illustrated embodiment of the CDER **310** includes data protection mechanisms such as a repository security system **404** to restrict access to the data, a data integrity system **408**, a data archival facility **410**, a backup and recovery system **412** and a repository accounting system **414**. Other data protection mechanisms may be used in conjunction with or in lieu of those shown in this embodiment, including any other data protection mechanisms which are commonly known in the art, such as, but in no way limited to, anti-virus scanners, data mirrors, failsafe systems that protect the access to the data in the event that the connected server is unable to communicate, sensing applications alone or in combination with a mote to monitor external or

internal conditions such as the ambient temperature, humidity or vibration, and emergency power systems such as an uninterruptible power supply.

The repository security system **404** may be included to enable the restriction of access to the data and data elements at multiple levels. The repository security system may include an authentication process to ensure that the user accessing or altering the data elements is a trusted source. The authentication process may include any such process commonly known in the art, including, but in no way limited to, password protection, the use of passcards, digital signatures or biometric authentication.

A data integrity system **408** may be used to further ensure that the reliability of the data is maintained. The data integrity system **408** protects data which may be accessed by two or more users or a single user with multiple sessions. The data integrity system **408** may be any such system that is known in the art, including, but not limited to, optimistic locking, pessimistic locking, locking on a column and application-level locking; all of which ensure safe distributed data editing. The data integrity system **408** may, for example, function so that if one user updates data while another user is working on it, neither user will be able to save any changes made without first viewing and acknowledging the changes that the other user has made.

The data archival system **410** may be used to ensure that all data is maintained and none is ever erased. This feature enables a designated user (*e.g.*, a system administrator) to "roll back" changes by users by reverting to a prior version or to view the database state at any point in the past. The data archival system **410** may be implemented, for example, by making a copy of any file that has been edited, incrementing the version associated with the data file and saving the copy in a data archive or storage facility. The data archival system may then be used to store all prior versions of the data. In alternate embodiments, the data archival system may be implemented by any other methodology commonly known in the art. It may also be configured wholly external from the CDER **310**.

The CDER may further include a repository accounting system **414**. The repository accounting system may be used to allow designated users (*e.g.*, a system administrator) to monitor access to the data and track all changes made to the same. Further, the repository accounting system **414** may compile and store the monitoring data, which may then be used to produce an audit trail of any modifications. The repository accounting system **414** may further be configured to limit or expand the quantum of information produced in the audit trail. The audit trail may include, for example, such information as the name of the user making the changes, the date and time such changes were made and other session related information, as well as a comparison chart summarizing how the data has been changed. The repository

accounting system **414** may be implemented in any way known in the art including, but not limited to, using an XQuery based web application.

Backup and recovery system **412** may also be employed to provide an added measure of security for the data stored in the CDER **310**. A backup and recovery system allows for high availability of the data and provides a means for the expedient and cost effective restoration of the data in the repository in the event of one of a number of catastrophic events, including a hardware or software failure and system infection with a virus.

In addition, the CDER **310** may further include a data transfer system **406**, which may be used to facilitate data transfer by means of import and export. The data transfer capability provides a mechanism for sharing large datasets between, for example, research institutions, which may enable a more thorough exploitation of the valuable dataset and thereby accelerate the search for improved treatment protocols. For example, a CDER built using an XML database (*e.g.*, Cerisent XQE) may import and export existing data elements both from and to data repositories, such as that of the National Cancer Institute Center for Bioinformation.

EXAMPLE 3

Form Tool

As illustrated in Figure 3, a form authoring tool **304** may be used in connection with the system of the present invention. Forms are used throughout the healthcare community for a variety of purposes, as will be readily appreciated by those of skill in the art. By way of example, forms are used to collect patient information in the course of a clinical study. The form authoring tool **304** provides for the expedient design of new forms that may be necessitated by, for example, a clinical trial. In addition, the form authoring tool **304** may be used to browse data elements stored in the CDER **310** that may then be selected for use on a form. As such, the form authoring tool **304** provides a designer with a means for building and editing forms with the common data elements stored in CDER **310**.

During the edit or design process, the form authoring tool **304** protects data integrity and permits form definitions to be changed independently of the data storage. As such, form elements can be added or removed while the application is in use without compromising data integrity. Once a form has been designed and/or editing is completed, a form definition corresponding to graphical representation of the form is generated and may then be saved in the processes database **104**. Of course, the form definition may alternatively be saved in a separate database such as a forms database (not shown). The form may then be rendered in various formats using a render agent **308**.

Figure 3 also illustrates a render agent **308** included with the common data element maintenance and electronic data capture system **300**. The render agent **308** may be used by the form authoring tool **304** to facilitate the design of new forms and editing of existing forms that may be used to collect varying subsets of patient data for multiple clinical trials. When designing a form, the render agent **308** reads data from the CDER **310**. The form authoring tool **304** may then be used to design a form. Once the form design is complete, the render agent **308** may be used to save the resulting form definition in the processes database **104** and to display the form in various formats. Thus, the render agent **308** may include any number of electronic data capture systems, such as a PhaseForward INFORM Electronic Data Capture system, the web form engine described below for electronic data capture via the Internet and/or a printer to render the form on paper.

Alternate embodiments of the present invention may include a web form engine. The web form engine may be used to electronically collect patient information (*e.g.*, patient medical history, updates on any symptoms the patient is experiencing, updates on any additional medications that the patient may be taking, updates to personal contact information, etc.). As a result, the web form engine may reduce the cost and time normally associated with manual data collection and entry. For example, in one embodiment, when a patient visits the website, the application generates a hypertext markup language ("HTML") form based on an abstract form definition. Once the patient has completed the form, the data on the form is collected and stored in an XML format.

Figure 5 depicts a block diagram of a system including the web form engine according to an embodiment of the present invention. The web form engine **500** may be used to design new forms useful in, for example, the collection of varying subsets of patient data for multiple clinical trials. When designing a form, the web form engine **500** reads data from the metadata repository **101**. This may, for example, permit the designer to select from a variety of categories of data elements which may be used on the newly designed form, such as patient's blood pressure, heart rate, white blood cell count, hemoglobin level, etc., depending on the nature of the clinical trial. Once the form design is complete, the web form engine **500** may be used to save the resulting form definition in the processes database **104**. Of course, the form definition may alternatively be saved in a separate database such as a forms database (not shown).

In addition, the web form engine **500** may be used to read abstract form definitions and facilitate electronic data collection. Once a form has been designed, it may be used to collect data from patients. The form engine **500** retrieves the form definition from the processes database **510**. The web form engine **500** may then construct an HTML representation of the

form based on its form definition, such that the form may be published by web server **506** and displayed for a user via output **112**, which may be a computer, a personal digital assistant ("PDA"), a mobile telephone or other similar device connected to the Internet or an intranet **202**. However, one skilled in the art will readily appreciate that the form may also be constructed and represented using other languages, including XML, SGML, VRML and the like.

In constructing the form, the web form engine **500** may use the data types of the common data elements to determine which control mechanisms will be used to display the data items. Such control mechanisms include, but are in no way limited to, radio buttons, drop down menus or lists and text fields. Once the form has been displayed, a user may then enter the requested data in the appropriate fields using input **110** (*e.g.*, a computer, PDA, mobile telephone, other device connected to the Internet or intranet **202**) and submit the form when completed. The web form engine **500** collects the input data and stores it in the facts database **102**.

Alternate embodiments of the web form engine **500** may include an archival system, wherein the web form engine **500** assigns a version number to each form stored in the facts database **102**. Each time a form is stored that matches the definition of an existing form in the facts database **102**, the form that exists in the database may be moved into a data archive (not shown) and the newer version put in its place in the facts database **102**. In doing so, researchers may conduct further research and track various trends in individual patients of a clinical trial, which may result in the development of improved treatment protocols.

In other embodiments, the web form engine **500** may further include various security features (not shown) to provide protection for the personal data that may be input by the user or displayed on a form generated by the web form engine. Such security features may include any authentication process commonly known in the art, including, but in no way limited to, password protection, the use of passcards, digital signatures or biometric authentication. In addition, the web form engine may include any data encryption scheme commonly known in the art, including, but in no way limited to, symmetric key encryption systems or public key encryption systems such as digital signatures, secure socket layer ("SSL"), transport layer security ("TLS") or any combination thereof, depending on the level of security deemed necessary and other considerations readily known to those of skill in the art.

Figure 6 illustrates a process in which an embodiment of the web form engine may be used to collect patient data via the Internet. First, a user may login at the website **602** from a remote computer and request a form **604**. Such website access may also be accomplished from a local computer within a clinic or a computer located within a clinic's intranet. Once the form has been requested, the web form engine retrieves the form definitions **608** from the metadata

repository and the corresponding factual data from the facts database and rules from the rules database, thereby allowing known responses to interrogatories (*e.g.*, the patient's name, address, prior medical history, status within treatment protocol, etc.) to be pre-populated for patient convenience. The web forms generator then generates the requested form **610**, which is displayed in HTML format on the user's computer. The user may complete the form **612** by inputting the requested data in the corresponding fields. When the user has completed the form, the data is encrypted **614** to protect private patient data, and transmitted to a web server within the clinic. The data is then decrypted by the web server, the markup is removed and the remaining data is stored in a flat file, which facilitates the data storage into an XML database. The web server forwards the data **616** to the web form engine, which archives **618** previous versions of the facts and stores the new version in the facts database.

EXAMPLE 4

Automated Clinical Trial Visit Scheduler

An embodiment of the present invention includes a visit scheduler. The visit scheduler facilitates the task of scheduling visits and procedures to be performed during patient visits, for example, as part of a clinical trial. Typically, this process is performed manually by staff members in a clinic or hospital, resulting in considerable time and expense. The clinic staff must fit each patient into the clinic schedule in light of what are oftentimes tight time constraints as to when the patient's visit can occur and the list of tests that must be run and the data gathered in accordance with the each patient's prescribed treatment protocol.

Figure 7 illustrates a block diagram of an embodiment of the present invention. As illustrated therein, a visit scheduler **700** may access the processes database **104** to obtain scheduling information and rules for a particular treatment regimen or protocol. The visit scheduler **700** accesses the facts database **102** and the rules database **103** to determine the patient type, the patient's current treatment regimen, the patient's status within the treatment regimen and any other appropriate information. This information may then be used to decide when the patient should be seen next. Further, the information may be used in conjunction with the processes database **104**, which may contain all data pertaining to the tasks to be performed during a patient visit. The processes database **104** may further identify the appropriate personnel that must be notified of those tasks during each patient visit within a particular treatment regimen.

The visit scheduler **700** may further include schedule authoring tool **708**, which allows the clinic or hospital to create new process definitions as new treatment protocols are

determined. When defining a new process, the visit scheduler **700** initiates the schedule authoring tool **708**, which may access the metadata repository **101** and the rules database **103** to define new metadata and rules for classifying and scheduling patients. The schedule authoring tool **708** then stores the newly defined rules in the rules database **103** and the newly defined metadata in the metadata repository **101**. However, the schedule authoring tool **708** may also choose existing rules to classify patients. In this embodiment of the invention, the schedule authoring tool **708** only writes data to the processes database **104**.

This visit scheduler **700** may be configured as a standalone application or may be embedded within a commercially available electronic mail, calendaring or productivity software package commonly known in the art (*e.g.*, Microsoft OUTLOOK). The visit scheduler **700** may also be constructed using any one or a combination of various programming languages as are commonly known in the art (*e.g.*, Visual Basic, Visual C++).

Figure 8 illustrates a process that may be used to schedule patient visits in a clinical trial and to define new data elements according to an embodiment of the present invention. First, a user may initiate a productivity application **802**. The user may then choose to initiate the visit scheduler **804**. Once the visit scheduler has been initiated, the user may define a new data element or schedule a clinical visit **806**. However, it should be noted that various other tasks may also be accomplished by the visit scheduler in alternate embodiments of the present invention, including, but in no way limited to, modifying and/or canceling visit schedules.

Where the user elects to define a new data element **808**, the user may define new metadata, rules or processes. If the user chooses to enter new metadata **810**, the entered data will be stored in the metadata repository **816**; if the user chooses to enter new rules **812**, the new rules will be stored in the rules database; and if the user opts to enter a new process **814**, the new process will be stored in the processes database **820**.

Alternatively, where the user elects to schedule a clinical visit **822**, the process data is retrieved **824** from the processes database. The user then enters patient data (if the patient is a new patient) or retrieves patient data **826** (if the patient is already a patient). Once the patient data has been entered or retrieved, the facts and rules data are retrieved **828** from the facts and rules databases, respectively. All available appointment slots are determined **830** in accordance with the patient treatment regimen and are presented to the user. The user may then select one of the available slots to create an appointment **832** and calendar it. If the user attempts to select an appointment that is outside the acceptable range defined by the user's treatment protocol, the visit scheduler warns the user and asks her to select another appointment. Once the user makes a selection, the appointment may then be entered into multiple calendars, including, but in no way

limited to, the patient's calendar and appropriate clinic staff's calendar(s) (*i.e.*, those that will be performing tasks associated with the patient's visit and/or laboratory technicians that will need to process and analyze specimens provided by the patient). Further, notifications may be distributed **834** to each of the aforementioned parties. The notifications may be delivered in a myriad of ways including but not limited to e-mail, facsimile, text or numeric paging, and voice messaging.

EXAMPLE 5

Electronic Medical Records Translator

An embodiment of the present invention includes an electronic medical records translator. The electronic medical records translator facilitates the rapid transformation of existing electronic medical records into a "fact" described in detail above. The electronic medical records translator may, for example, read the metadata in the metadata repository **101** in conjunction with the rules in the rules database **103** to determine how to properly translate the metadata into a fact. Once the translation is complete the resulting fact may then be stored in the facts database **102**. However, if the electronic medical record includes data for which there is no rule with which the data may be translated, the data may nevertheless be stored in the metadata repository **101** and flagged and later reviewed using an editing tool (*e.g.*, the metadata curation tool described above).

While the description above refers to particular embodiments of the present invention, it will be understood that many alternatives, modifications and variations may be made without departing from the spirit thereof. The accompanying claims are intended to embrace such alternatives, modifications and variations as would fall within the true scope and spirit of the present invention. The presently disclosed embodiments are therefore to be considered in all respects as illustrative and not restrictive, the scope of the invention being indicated by the appended claims, rather than the foregoing description, and all changes which come within the meaning and range of equivalency of the claims are therefore intended to be embraced therein.